# Co-clustering de données textuelles et continues

Margot Selosse, Julien Jacques, Christophe Biernacki

# Co-clustering de données textuelles et continues

Margot Selosse [1] & Julien Jacques [2] & Christophe Biernacki [3]

[1] *Université de Lyon, Lyon 2, ERIC EA 3083 margot.selosse@univ-lyon2.fr*
[2] *Université de Lyon, Lyon 2, ERIC EA 3083 julien.jacques@univ-lyon2.fr*
[3] *Inria, Université de Lille, CNRS, christophe.biernacki@inria.fr*

**Résumé.** Le clustering est un outil essentiel pour l'analyse de données. C'est une manière de résumer un jeu de données en formant des groupes homogènes d'observations (les clusters). Cependant, le phénomène «big-data »a fait croître le nombre de variables, conduisant à l'émergence de jeux de données de grande dimension, parfois à un niveau tel que les techniques de clustering ne sont plus toujours adaptées pour discerner des structures. En effet, l'analyse d'un cluster repose généralement sur un représentant de ce cluster (par exemple la moyenne). Néanmoins, ce dernier est lui-même décrit par un grand nombre de variables, ce qui rend difficile l'interprétation et hasardeuse l'estimation. De cette observation naît le besoin de «résumer »aussi les variables, ce que leur regroupement en clusters peut permettre, de façon symmétrique au regroupement classique des individus en clusters. Le co-clustering est alors une méthode candidate car elle réalise un clustering simultané des lignes et des colonnes. Dans le cas de l'analyse de données textuelles, et notamment le clustering de document, le co-clustering est un thème largement étudié lors de ces dernières années. Cependant, la plupart des approches ne permettent pas de prendre en compte, en plus des données textuelles, d'autres variables. Le travail présenté propose une extension du modèle des blocs latents pour des jeux de données avec des variables textuelles et continues.

**Keywords.** Co-clustering, données hétérogènes, modèle des blocs latents

**Abstract.** Clustering is an essential tool in data analysis. It is a way of summarizing a dataset by forming homogeneous groups of observations (clusters). However, the "big-data" phenomenon has enlarged the number of features, and increased the emergence of high-dimensional datasets, sometimes in such an extent that clustering techniques are therefore not always adapted to discern structures. Indeed, the analysis of a cluster usually relies on a representative of the cluster (e.g: mean). Yet, this latter is itself described by a high number of features, which makes it difficult to interpret. From this comes the need to also «summarize »the features. Co-clustering methods is a good candidate for performing this task because it realizes a joint clustering of rows and columns. Text-mining, and document clustering has been being well studied through the last years. However, most of the approaches can not take into account variables that are not textual. This work suggests an extension of the Latent Block Model for datasets made of textual and quantitative data.

**Keywords.** Co-clustering, heterogeneous data, latent bloc model

# 1 Introduction

## 1.1 Les TED Talks

Le jeu de données utilisé dans cette application contient les informations de 2467 TED talks[1]. Les TED talks sont des conférences dans lesquelles des intervenants prennent la parole sur un sujet (high-tech, psychologie, sport...). Nous nous sommes intéressés au script de ces interventions, qui sont disponibles sous forme de texte, ainsi qu'à leur note qui a été attribuée par les spectateurs. Le système de note est particulier pour les TED talks. Quatorze adjectifs pouvant qualifier un talk sont définis : Inspiring, OK, Beautiful, Ingenious, Courageous, Unconvincing, Persuasive, Obnoxious, Jaw-dropping, Fascinating, Informative, Funny, Confusing, Longwinded. Un spectateur peut alors cliquer sur les adjectifs qui lui paraissent correspondre à l'intervention qu'il vient de voir. Après un prétraitement, notre jeu de données est donc vu comme deux matrices $\boldsymbol{x}^{(1)}$ et $\boldsymbol{x}^{(2)}$ côte à côte. La matrice $\boldsymbol{x}^{(1)}$, de dimension $2467 \times 40137$ est la matrice Documents-Termes des scripts de chaque TED talk, où un élément dénombre les occurences d'un terme dans le document. Nous proposons de modéliser ce nombre d'occurences par une loi de Poisson. La matrice $\boldsymbol{x}^{(2)}$, de dimension $2467 \times 14$ représente, pour chaque talk, le nombre de spectateurs ayant voté pour chacun des adjectifs parmi la liste citée. Etant donné le grand nombre de votes, nous proposons de modéliser ce nombre par une loi normale.

Le premier objectif de ce travail est d'effectuer un clustering des TED talks, à partir de leur données textuelles, c'est à dire la matrice Documents-Termes, mais également à partir des notes des spectateurs. Le co-clustering est une technique qui réalise un clustering simultané des lignes et des colonnes. Cela fait donc apparaître des blocs, qui sont le croisement d'un cluster en ligne et d'un cluster en colonne. Dans notre cadre, le co-clustering sert donc non seulement à créer des clusters de documents mais aussi des clusters de mots ce qui est un avantage important, par exemple, pour repérer les termes caractéristiques d'un groupe de documents. De plus, ajouter l'information liée aux notes des spectateurs permet de mieux interpréter les résultats, et d'améliorer la synthèse du jeu de données. Cependant, les notes ne peuvent pas être incluses au même titre que les termes dans la matrice de données car elles sont de nature différente. Il faut donc réaliser un co-clustering de données mixtes, c'est-à-dire, avec des variables qui ne sont pas de même nature.

## 1.2 Notations

Nous considérons la matrice, avec deux différents types de variables, $N$ lignes et $J = J_1 + J_2$ colonnes. Pour représenter ce jeu de données, les variables qui ne sont pas de même nature sont séparées. Le résultat correspond donc à deux matrices de $N$ lignes et $J_1, J_2$ colonnes, respectivement.

[1] `https://www.kaggle.com/rounakbanik/ted-talks/data`

$$\boldsymbol{x} = \left[\left[ \quad \boldsymbol{x}^{(1)} \quad \right] \left[ \quad \boldsymbol{x}^{(2)} \quad \right]\right]$$

Nous notons $x_{ij}^{(d)}$ l'élément de la $i$-ème ligne et $j$-ème colonne de la $d$-ème matrice avec $d \in \{1, 2\}$. Etant dans un contexte de co-clustering, nous supposons qu'il existe $G$ clusters en ligne et $H = H_1 + H_2$ clusters en colonne.

# 2 Modèle des blocs latents multiples

## 2.1 Hypothèses du modèle

Le modèle des blocs latents [2] est l'un des modèles statistiques les plus utilisés en co-clustering. Il repose sur l'hypothèse que les élements au sein d'un bloc sont les réalisations de variables aléatoires qui suivent une distribution paramétrique, spécifique au bloc. L'avantage de cette approche est que ces paramètres permettent d'interpréter les blocs résultants. De plus, elle permet de produire une estimation très parcimonieuse, même en grande dimension. Néanmoins, cette méthode ne peut pas être utilisée dans la cas de données de types différents. Effectivement, si les éléments d'un bloc ne sont pas même nature, il n'est pas possible de considérer qu'ils aient été tirés selon une même distribution. Le modèle des blocs latents multiples, introduit dans [4] et généralisé pour ce travail, offre une manière de pallier ce problème. Le co-clustering est réalisé de telle manière que les variables de différents types ne peuvent pas faire partie d'un même cluster en colonne.

Nous supposons que pour chaque matrice $\boldsymbol{x}^{(d)}$ $(1 \leq d \leq 2)$, il existe une partition des lignes $\boldsymbol{v} = (v_{ig})_{i,g}$ et une partition des colonnes $\boldsymbol{w}^d = (w_{jh}^{(d)})_{j,h}$. Les plages de variation de $i, j, g, h$ et $d$ sont omises pour faciliter l'écriture. $v_{ig}$ est égal à 1 si la ligne $i$ appartient au cluster $g$, et 0 sinon. Parallèlement, $w_{jh}^{(d)}$ est égal à 1 lorsque la colonne $j$ appartient au cluster $h$, et 0 autrement. Dans notre cas, la matrice $\boldsymbol{x}$ est vue comme deux matrices de différente nature, comme décrit en partie 1.2. Le modèle des blocs latents multiples repose sur l'hypothèse suivante :

$$p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} | \boldsymbol{v}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}) = p(\boldsymbol{x}^{(1)} | \boldsymbol{v}, \boldsymbol{w}^{(1)}) \times p(\boldsymbol{x}^{(2)} | \boldsymbol{v}, \boldsymbol{w}^{(2)}).$$

De plus, les variables aléatoires univariées $x_{ij}^{(d)}$ sont supposées être indépendantes conditionnellement aux partitions $\boldsymbol{v}$ et $\boldsymbol{w}^{(d)}$. Ainsi, la fonction de densité de probabilité de $\boldsymbol{x}$ conditionnellement à $\boldsymbol{v}$, $(\boldsymbol{w}^{(1)})$ et $(\boldsymbol{w}^{(2)})$ peut être écrite :

$$p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} | \boldsymbol{v}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}; \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}) \;=\; \prod_{i,j,g,h,d} p(x_{ij}^{(d)}; \alpha_{gh}^{(d)})^{v_{ig} w_{jh}^{(d)}},$$

où $\boldsymbol{\alpha}^{(d)} = (\alpha_{gh}^{(d)})_{g,h}$ sont les paramètres de la distribution du bloc $(g, h)$ de la matrice $\boldsymbol{x}^{(d)}$. Enfin, les variables latentes $\boldsymbol{v}$, $\boldsymbol{w}^{(1)}$ et $\boldsymbol{w}^{(2)}$ sont considérées indépendantes, c'est pourquoi $p(\boldsymbol{v}, \boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}; \boldsymbol{\pi}, \boldsymbol{\rho}^{(1)}, \boldsymbol{\rho}^{(2)}) \;=\; p(\boldsymbol{v}; \boldsymbol{\pi}) p(\boldsymbol{w}^{(1)}; \boldsymbol{\rho}^{(1)}) p(\boldsymbol{w}^{(2)}; \boldsymbol{\rho}^{(2)})$ avec $\boldsymbol{\pi} = (\pi_g)_g$ les proportions du mélange en ligne et $\boldsymbol{\rho} = (\rho_h^{(d)})_h$ les proportions du mélange en colonne :

$$p(\boldsymbol{v};\boldsymbol{\pi}) \;=\; \prod_{i,g}\pi_g^{v_{ig}} \;,\; p(\boldsymbol{w}^{(d)};\boldsymbol{\rho}^{(d)}) \;=\; \prod_{j,h}{\rho_h^{(d)}}^{w_{jh}^{(d)}},\ \pi_g = p(v_{ig}=1) \text{ et } \rho_h^{(d)} = p(w_{jh}^{(d)}=1).$$

Ainsi, avec $V$ l'ensemble des $\boldsymbol{v}$ possibles, et $W^{(d)}$ l'ensemble des $\boldsymbol{w}^{(d)}$ possibles pour $1 \leq d \leq 2$, la densité de probabilité de $\boldsymbol{x}$ est définie par :

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{(v,w^{(1)},w^{(2)})\in V\times W^{(1)}\times W^{(2)}} \prod_{i,g}\pi_g^{v_{ig}} \prod_{d,j,h}{\rho_h^{(d)}}^{w_{jh}^{(d)}} \prod_{i,j,g,d,h} p(x_{ij}^{(d)};\alpha_{gh}^{(d)})^{v_{ig}w_{jh}^{(d)}},$$

avec $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}^{(1)}, \boldsymbol{\rho}^{(2)}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)})$, l'ensemble des paramètres du modèle.

## 2.2 Modélisation des données

Pour les données textuelles, il est supposé que pour chaque bloc $(g,h)$, les variables $x_{ij}^{(1)}$ suivent une distribution de Poisson $\mathcal{P}(\mu_i\nu_j\gamma_{gh})$. Le paramètre de Poisson est défini par $\mu_i$ et $\nu_j$, respectivement l'effet de la ligne $i$ et de la colonne $j$, et $\gamma_{gh}$, l'effet du bloc $(g,h)$ [2]. La probabilité conditionnelle s'écrit alors :

$$p(x_{ij}^{(1)}|v_{ig}=1, w_{jh}^{(1)}=1) = \tfrac{1}{x_{ij}^{(1)}!}e^{-\mu_i\nu_j\gamma_{gh}}(\mu_i\nu_j\gamma_{gh})^{x_{ij}^{(1)}}.$$

Pour les notes, la distribution Gaussienne est utilisée. La probabilité conditionnelle est donc écrite :

$$p(x_{ij}^{(2)}|v_{ig}=1, w_{jh}^{(2)}=1) = \tfrac{1}{\sqrt{2\pi\sigma_{gh}^2}}\exp\{\tfrac{-1}{2\sigma_{gh}^2}(x_{ij}^{(2)}-\mu_{gh})^2\}.$$

Nous avons donc $\boldsymbol{\alpha}^{(1)} = (\boldsymbol{\gamma})$ et $\boldsymbol{\alpha}^{(1)} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, avec $\boldsymbol{\gamma} = (\gamma_{gh})_{g,h}$, $\boldsymbol{\mu} = (\mu_{gh})_{g,h}$ et $\boldsymbol{\sigma}^2 = (\sigma_{gh}^2)_{g,h}$.

# 3 Inférence du modèle avec l'algorithme SEM-Gibbs

L'algorithme EM [1] est une méthode classique pour maximiser la vraisemblance en présence de variables latentes. Cependant, dans le cadre du co-clustering, il est impossible de l'utiliser car le calcul est trop coûteux. Il existe plusieurs alternatives. Nous avons choisi d'utiliser ici un EM stochastique couplé à un échantillonneur de Gibbs (SEM-Gibbs). Cette méthode a prouvé son efficacité [3]. L'algorithme SEM-Gibbs itère un certain nombre de fois $Q$ les étapes suivantes ($(q)$ représentant la $q$-ieme itération) :

**1. Echantillonnage en ligne.** Générer $\boldsymbol{v}^{(q)}$ selon
$\log(p(v_{ig}^{(q)} = 1 \mid \boldsymbol{x}, \boldsymbol{w}^{(1)^{(q-1)}}, \boldsymbol{w}^{(2)^{(q-1)}})) = A(i,g) + K(i)$, avec :
$A(i,g) = \log(\pi_g^{(q-1)}) + \sum_{j,h} w_{jh}^{(1)^{(q-1)}}(-\mu_i\nu_j\gamma_{gh}^{(q-1)} + x_{ij}^{(1)}\log(\mu_i\nu_j\gamma_{gh}) - \log(x_{ij}^{(1)}!))$
$+ \sum_{j,h} w_{jh}^{(2)}\big(\tfrac{-1}{2\sigma_{gh}^{2\ (q-1)}}(x_{ij}^{(2)} - \mu_{gh}^{(q-1)})^2 - \tfrac{1}{2}\log(2\pi\sigma_{gh}^{2\ (q-1)})\big)$.
Ici, $K(i) = -\sum_g A(i,g)$.

**2. Actualisation des paramètres.** Cette étape met à jour les paramètres $\boldsymbol{\theta}^{(q)}$ pour maximiser la log-vraisemblance complétée. Les proportions de mélange sont actualisées par : $\pi_g^{(q)} = \frac{1}{N}\sum_i v_{ig}^{(q)}$, et $\rho_h^{(d)(q)} = \frac{1}{H_d}\sum_j w_{jh}^{(d)(q-1)}$. De plus, $\boldsymbol{\alpha}^{(d)(q)}$ est aussi mis à jour, les calculs dépendant du type des variables de $\boldsymbol{x}^{(d)}$.
Pour les données de comptage :

$$\gamma_{gh}^{(q)} = \frac{1}{n_{g.}n_{.h}}\sum_{i,j} v_{ig}^{(q)} w_{jh}^{(1)(q-1)} x_{ij}^{(1)},\ \mu_i^{(q)} = n_{i.},\ \nu_j^{(q)} = n_{.j}.$$

Ici, $n_{g.} = \sum_{i,j} v_{ig}^{(q)} x_{ij}^{(1)}$, $n_{.h} = \sum_{i,j} w_{jh}^{(1)(q-1)} x_{ij}^{(1)}$, $n_{i.} = \sum_j x_{ij}^{(1)}$ et $n_{.j} = \sum_i x_{ij}^{(1)}$.
Pour les données continues :

$$\mu_{gh}^{(q)} = \frac{1}{N_{gh}}\sum_{i,j} v_{ig}^{(q)} w_{jh}^{(2)(q-1)} x_{ij}^{(2)},\ \sigma_{gh}^{2\ (q)} = \frac{1}{N_{gh}-1}\sum_{i,j} v_{ig}^{(q)} w_{jh}^{(2)(q-1)} (x_{ij}^{(2)} - \mu_{gh}^{(q)})^2.$$

Ici, $N_{gh} = \sum_{i,j,g,h} v_{ig}^{(q)} w_{jh}^{(2)(q)}$ représente le nombre d'éléments présents dans le bloc $(g,h)$.

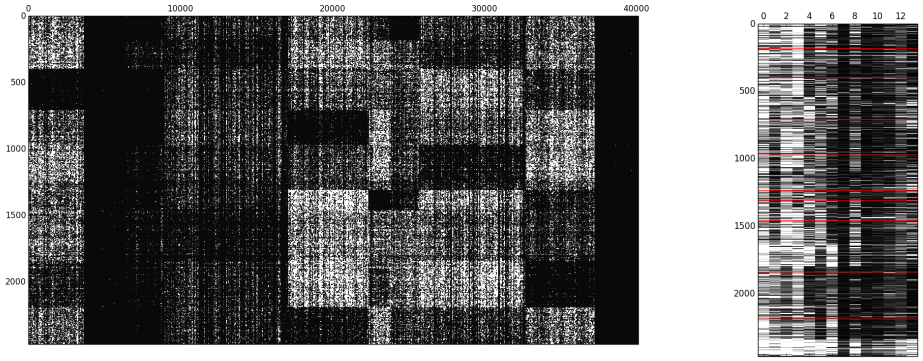
**3. Echantillonnage en colonne.** Pour $d \in \{1,2\}$, générer les partitions en colonnes $\boldsymbol{w}^{(d)(q)}$ de la $d$-ieme table $\boldsymbol{x}^{(d)}$ avec :
$\log(p(w_{jh}^{(1)(q)} = 1 \mid \boldsymbol{x}^{(1)}, \boldsymbol{v}^{(q)})) = B_1(j,h) + L_1(j)$ et
$B_1(j,h) = \log \rho_h^{(1)(q)} + \sum_{i,g} v_{ig}^{(q)}(-\mu_i\nu_j\gamma_{gh}^{(q-1)} + x_{ij}^{(1)}\log(\mu_i\nu_j\gamma_{gh}) - \log(x_{ij}^{(1)}!))$,
$\log(p(w_{jh}^{(2)(q)} = 1 \mid \boldsymbol{x}^{(2)}, \boldsymbol{v}^{(q)})) = B_2(j,h) + L_2(j)$ et
$B_2(j,h) = \log \rho_h^{(2)(q)} + \sum_{i,g} v_{ig}^{(q)}\big(\frac{-1}{2\sigma_{gh}^{2\ (q)}}(x_{ij}^{(2)} - \mu_{gh}^{(q)})^2 - \frac{1}{2}\log(2\pi\sigma_{gh}^{2\ (q)})\big)$.
Ici, $L_1(j) = -\sum_h B_1(j,h)$ et $L_2(j) = -\sum_h B_2(j,h)$.

**4. Actualisation des paramètres.** Cette étape met de nouveau à jour $\boldsymbol{\theta}^{(q)}$, comme dans l'étape 2, mais avec les partitions $\boldsymbol{v}$, $\boldsymbol{w}^{(1)(q)}$ $\boldsymbol{w}^{(2)(q)}$.

A la fin des itérations, $\hat{\boldsymbol{\theta}}$ est obtenu en moyennant les valeurs de $\boldsymbol{\theta}^{(q)}$ obtenue au cours des itérations (en enlevant une période de chauffe). Les partitions $\boldsymbol{v}$, $\boldsymbol{w}^{(1)}$ et $\boldsymbol{w}^{(2)}$ sont ensuite échantillonnées avec $\hat{\boldsymbol{\theta}}$.

# 4 Résultats sur les TED talks

Nous avons fixé arbitrairement les nombres de clusters à $(G = 10, H_1 = 10, H_2 = 2)$. Le clustering en colonne des notes fait apparaître deux groupes distincts d'adjectifs, qui




(a) (b)

Figure 1: Co-clustering de la matrice de données. 1a représente la matrice Documents-Termes et 1b représente la matrice de notes attribuées par les spectateurs.

peuvent être vus comme le groupe péjoratifs (Confusing, Obnoxious, Longwinded, Unconvincing, OK) et le groupe mélioratif (les autres). La Figure 1a représente la matrice Documents-Termes et la Figure 1b la matrice de notes, ordonnées selon les partitions en ligne, qui leur sont communes, et leurs clusters en colonne respectifs. Le co-clustering est riche en information. Par exemple, nous remarquons que le cluster de documents qui a le taux de votes pour les adjectifs mélioratifs le plus élevé est très lié à la politique : il contient notamment des talks prônant la démocratie et la non-violence. Par exemple quelques titres de talks de ce cluster sont : «Why we need to end the War on Drugs», «Social experiments to fight poverty», «How racism makes us sick».

# References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *JRSS, series B*, 39(1):1–38, 1977.

[2] G. Govaert and M. Nadif. *Co-Clustering*. ISTE Ltd and John Wiley & Sons Inc., 2014.

[3] C. Keribin, G. Govaert, and G. Celeux. Estimation d'un modèle à blocs latents par l'algorithme SEM. In *42èmes Journées de Statistique*, Marseille, France, 2010.

[4] V. Robert. *Classification croisée pour l'analyse de bases de données de grandes dimensions de pharmacovigilance*. PhD thesis, 2017.